

SEMESTER END EXAMINATION, APRIL-MAY, 2025

Course Name: - B.Tech

Semester:- 6th

Paper Name: - Data Mining and Warehousing

Paper Code:- TCS 063

Time - 3 Hrs + 20 minutes per hour extra time for V.I. & examinees with writer.

Max Marks-70

Additional 30 Minutes for Mid-Test.

समय- 3 घण्टे + 20 मिनट प्रति घंटे अतिरिक्त-दृष्टिबाधित एवं सह लेखक परीक्षार्थियों के लिए।
30 मिनट अतिरिक्त मिड-टेस्ट के लिए।

अधिकतम अंक-70

Instructions:

- The question paper consists of three sections namely A, B, C. All sections are compulsory.
- Section A- Each question carries 3 mark. All questions are compulsory.
- Section B- Answer any 5 out of 7 given questions. Each question carries 7 marks.
- Section C- Answer any 2 out of 3 given questions. Each question carries 10 marks.
- Section D- Each question carries 02 mark. All questions are compulsory.

निर्देश:

- प्रश्न पत्र में तीन खण्ड अ, ब, व स हैं। सभी खण्ड अनिवार्य हैं।
- खण्ड-अ में प्रत्येक प्रश्न तीन अंक का है। सभी प्रश्न अनिवार्य हैं।
- खण्ड-ब में सात प्रश्नों में से किन्हीं पाँच प्रश्नों के उत्तर दें। प्रत्येक प्रश्न सात अंक का है।
- खण्ड-स में तीन प्रश्नों में से किन्हीं दो प्रश्नों के उत्तर दें। प्रत्येक प्रश्न 10 अंक का है।
- खण्ड-द में प्रत्येक प्रश्न 02 अंक का है। सभी प्रश्न अनिवार्य हैं।

Section - A खण्ड-अ

Objective Questions (वस्तुनिष्ठप्रश्न)

1. Answer all the following questions.

5×3=15

निम्नलिखित सभी प्रश्न अनिवार्य हैं।

1- Association rule mining is most suitable for which of the following applications?

- A) Predicting a customer's credit risk
- B) Detecting anomalies in network traffic
- C) Market Basket Analysis
- D) Clustering web users by interest

2- Which of the following OLAP operations can indirectly perform the function of slicing?

- A) Dice
- B) Drill-down
- C) Roll-up
- D) Pivot

3- Which of the following techniques can handle non-linear decision boundaries effectively?

- A) Decision trees
- B) Naive Bayes
- C) Linear regression
- D) K-Nearest Neighbors (KNN)

4- You are analyzing a dataset of people's heights. One value is 35 cm. Based on context, this is likely an outlier due to:

- A) Data skew
- B) Normal variance
- C) Wrong Input
- D) Shorter population

5- Which of the following algorithms is most commonly applied for classification, rather than clustering or association rule mining?

- A) K-means
- B) Decision Tree
- C) Apriori
- D) FP-Growth

Section - B खण्ड-ब

Short Answer Questions (लघुउत्तरीय प्रश्न)

2. Answer any five of the following questions.

5×7=35

निम्नलिखित में से किन्हीं 5 प्रश्नों के उत्तर दें।

- (i) What is ETL (Extract, Transform, Load) in the context of data warehousing? Discuss its importance and the challenges involved in the ETL process?
- (ii) Discuss the advantages and disadvantages of MOLAP over ROLAP in terms of data storage, query performance, and real-time analysis. In what scenarios would ROLAP be preferred over MOLAP?
- (iii) Distinguish between OLAP and OLTP systems ?
- (iv) Elaborate on the Apriori algorithm and describe how it employs the concepts of support and confidence in the generation of association rules.?
- (v) Describe the top-down and bottom-up approaches to Data Warehouse Architecture. Provide a diagrammatic representation of both approaches and explain each approach in detail.?
- (vi) Explain the significance of Outlier and Noise Analysis in Data Mining. Describe different techniques for detecting outliers, How do outliers differ from noise, and how can detecting them help identify anomalies like fraud in large datasets?
- (vii) Discuss the advantages and disadvantages of using a Data Mart compared to a full-scale Data Warehouse.

Section - C खण्ड-स

Descriptive Questions (विवरणात्मकप्रश्न)

3. Answer any 02 question of the following question.

2×10=20

निम्नलिखित में से किन्हीं 2 प्रश्न के उत्तर दें।

- (i) Describe the key features of OLAP systems, including slice, dice, drill-down, and roll-up operations. Explain how these features are used to analyse multi-dimensional data and improve decision-making. Include examples of business scenarios where each feature would be applied.?
- (ii) In the KDD process, how does the data preprocessing stage impact the quality of the final knowledge discovered? Explain how data cleaning, integration, and transformation techniques help address the challenges of noisy, missing, and inconsistent data.?

(iii) Scenario:

A retail company, ShopCart, wants to analyze customer buying patterns using Market Basket Analysis to find out which items are frequently bought together. Their sales data comes from different sources like physical stores and an online shopping app. To manage this scattered data effectively and support this analysis, the company plans to set up a Data Warehouse.

Question:

You're assisting ShopCart in building a Data Warehouse to perform Market Basket Analysis. Briefly explain how the staging area, data integration layer, and presentation layer work together to collect, clean, transform, and prepare data for analysis. How does this layered architecture support the discovery of product association rules?

SEMESTER END EXAMINATION, APRIL-MAY, 2025

Mid-Test

Course Name: - B.Tech

Paper Name: - Data Mining and Warehousing

Time - 30 minutes.

Semester:- 6th

Paper Code:- TCS 063

Max Marks-20

All questions are compulsory.

2×10=20

सभी प्रश्न अनिवार्य हैं।

Objective Questions.

बहुविकल्पीय प्रश्न।

1- In the context of data integration, what issue can lead to "data redundancy"?

- A) Conflicting data values
- B) Presence of foreign keys
- C) Use of ETL tools
- D) Different naming conventions in schema

2- A rule {Milk} → {Bread} has support = 0.4 and confidence = 0.8. If there are 1000 transactions, how many contain both Milk and Bread?

- A) 320
- B) 400
- C) 800
- D) 500

3- What is a true statement about outlier analysis in a dataset?

- A) Outliers are always considered errors and must be removed
- B) Outliers are rare and uninteresting occurrences in data
- C) Outlier analysis helps uncover anomalies that may be critical for decision making
- D) Outlier detection only applies to time-series datasets

4- What is a "slice" operation in OLAP?

- A) Merging two dimensions into one
- B) Fixing a value for one dimension and retrieving a sub-cube
- C) Splitting data into smaller chunks for parallel processing
- D) Normalizing data in a cube

5- A data warehouse has 4 dimensions. If each has a 3-level hierarchy, how many total concept hierarchy levels are there in the warehouse?

- A) 4
- B) 12
- C) 7
- D) 9

6- What is one of the major challenges of using decision trees for classification?

- A) They are very slow to train.
- B) They are prone to overfitting, especially with deep trees.
- C) They cannot handle categorical data.
- D) They require a large amount of memory to store.

7- If a dataset contains several overlapping clusters with arbitrary shapes, which clustering algorithm is most appropriate?

- A) K-Means
- B) DBSCAN
- C) K-Medoids
- D) Agglomerative Clustering

8- Which of the following scenarios is not typically an outlier case?

- A) A transaction 100× bigger than others
- B) A data point falling on the cluster boundary
- C) A lone point in a sparse region of DBSCAN
- D) A student scoring 99 in a test where average is 50

9- Which feature is common in OLAP but not in OLTP?

- A) Real-time updates
- B) High normalization
- C) Drill-down and roll-up
- D) Transactional integrity

10- A Data cube has 3 dimensions: A(3 values), B(4 values), C(2 values). How many total base cells are there?

- A) 24
- B) 9
- C) 12
- D) 36
